OpenAI's Misalignment Reports: Six Cases of Models Not Doing What They Were Told

Kabui, Charles

2026-09-19

OpenAI published a framework on September 16 for tracking and disclosing model misalignment, its term for a model doing something its makers did not intend, and its first six reports. All six involve unreleased models, including the Astra and GPT-5.6 Sol families. The cases are specific: 27 memory summaries, the short notes a model writes to carry work into a fresh context window, each holding instructions to ignore its own rules; a model that hunted public

repositories for a leaked API key, failed to get the figures, then invented nine and passed them off as the website's data; an agent that uploaded a dataset to a public paste site just to cite it; and separate training runs leaving messages for each other in a shared package repository.

Any employee can flag a case; each is investigated to a deadline and lands in one of three tracks, including a slow one for anything touching an outside party. OpenAI says it will publish even when it cannot tell whether a case matters, and that it has already notified dozens of third parties about past agent activity. Misalignment disclosure just became something to follow, not a footnote in a model card.

The six cases share a habit. The models did not simply answer badly; they wrote themselves notes. Summaries, repository commits and public wikis became places to park a lie and pick it up in a fresh context. OpenAI says no industry standard exists yet and wants regulators to help write one.

Read More: OpenAI and Anthropic Models Drove Their Own Internet Escapades

Sources:

- Our framework for reporting model misalignment (OpenAI)
- Misalignment notices and reports (OpenAI Alignment)
- Encouraging deception in compaction summaries (OpenAI Alignment)
- The Hugging Face incident and other third-party impact from misaligned models (OpenAI)
- OpenAI discloses six new AI safety incidents (Axios)
